Avaliação de Pequenos Disjuntos Utilizando Medidas de Precisão de Regras

Alan Keller Gomes[1] [2]

[1] Centro Federal de Educação Tecnológica de Urutai – CEFET Urutai
Fazenda Palmital, Km 2, Zona Rural, CEP 75790-000, Urutai, GO
www.cefeturutai.edu.br — alankeller@uol.com.br
[2] Faculdade Alves Faria – ALFA — Av. Perimetral Norte, 4129
Vila João Vaz, CEP 74445-190, Goiânia, GO
www.alfa.br — alankeller@alfa.br

Resumo O conhecimento induzido por sistemas de aprendizado de máquina simbólico, a partir de um conjunto de exemplos, pode ser descrito como um conjunto de regras *if-then*, ou seja, regras na forma $B \rightarrow H$. A base padrão para a avaliação dessas regras é a tabela de contingência, da qual podem ser obtidos os valores absolutos (cardinalidades). A partir desses valores, medidas de avaliação como cobertura e precisão podem ser calculadas. Pequenos disjuntos são regras que cobrem um pequeno número de exemplos. Mesmo sendo propensos ao erro e com pouca generalidade, o grau de surpresa (interessabilidade) desse conhecimento pode ser calculado. Este trabalho tem por objetivo ressaltar a relação entre cardinalidades e medidas de precisão, na avaliação de de regras selecionadas pelo grau de surpresa de pequenos disjuntos, com o propósito de identificar aspectos referentes à cobertura e à precisão de classificação de exemplos por parte desse conhecimento.

1 Introdução

Sistemas de Aprendizado de Máquina (AM) supervisionado são capazes de induzir uma hipótese (ou classificador), descrito na forma de árvores de decisão ou regras de classificação, a partir de um conjunto de dados previamente rotulados.

Um classificador, ou hipótese **h**, geralmente pode ser transformado em um conjunto de regras *if-then*, ou seja, regras do tipo *Corpo* $\rightarrow$ *Cabeça* ou $Body \rightarrow Head$. Uma regra R_u pode então ser resumidamente denotada como $B \rightarrow H$. Assim, **h** consiste de um conjunto de NR regras $R_u, u = 1, ..., NR$, ou seja, $\mathbf{h} = \{R_1, ..., R_{NR}\}$, ao qual denominamos *classificador simbólico*.

Em tarefas voltadas para a descoberta de conhecimento, deseja-se que as regras descobertas sejam mais genéricas e precisas tanto quanto possível. A utilização de medidas de avaliação, como as propostas por Lavrac em [1], tem o intuito de destacar regras que apresentam balanceamento entre generalização e precisão.

Sob o ponto de vista da precisão de classificação, *pequenos disjuntos* são regras que cobrem um número pequeno de exemplos e, em geral, são indesejáveis porque

apresentam pouca generalidade e são propensos ao erro. No entanto, sob o ponto de vista do grau de surpresa, eles tem o potencial para capturar conhecimento inesperado a partir dos dados [2].

Utilizando medidas de interessabilidade é possível selecionar o conhecimento mais interessante, inesperado ou surpreendente para o usuário [5][3]. No entanto, esse usuário também pode estar interessado em observar algum aspecto relacionado à cobertura e à precisão na classificação de exemplos. Nesse sentido, este trabalho tem por objetivo observar a relação entre cardinalidades e medidas de precisão de regras a partir da avaliação de pequenos disjuntos (conhecimento interessante).

O $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ é um sistema computacional protótipo, desenvolvido na linguagem de programação lógica Prolog [3], que implementa funcionalidades voltadas para AM, tais como a análise automática e a combinação de regras na forma $B \rightarrow H$. O Módulo de Análise de Regras (MAR) do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ [4], fazendo uso de medidas de avaliação e de interessabilidade, fornece subsídios para a análise automática dessas regras.

Neste trabalho, o MAR do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ é utilizado, primeiramente, para calcular o grau de surpresa de pequenos disjuntos, induzidos a partir de diferentes conjuntos de exemplos. Em seguinda, avaliar esse conhecimento utilizando medidas de precisão de regras.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a avaliação de regras, medida do grau de surpresa de pequenos disjuntos e medidas de precisão de regras; a Seção 3 apresenta a construção do experimento onde pequenos disjuntos, induzidos a partir de diferentes conjuntos de dados, são avaliados utilizando medidas implementadas no MAR do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$; a Seção 5 apresenta os resultados obtidos e, a Seção 6, as considerações finais referentes a este trabalho.

2 Avaliação de Regras e Medidas Utilizadas

A avaliação de cada regra R_u na forma $B \rightarrow H$, do conjunto de regras que constituem $\mathbf{h}$, é realizada utilizando a *tabela de contingência* da regra, ilustrada na Tabela 1. Essa tabela é calculada utilizando um conjunto de exemplos do domínio $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, onde $\mathbf{x}_i$ é um vetor de valores dos atributos do exemplo i e $y_i \in \{C_1, C_2, ..., C_{NCl}\}$ a sua classe. NCl é o número de classes do conjunto de exemplos. A tabela de contingência é uma generalização da *matriz de confusão*, utilizada como base para computar medidas de avaliação de hipóteses.

Na Tabela 1, B denota o conjunto de exemplos para os quais o corpo da regra é verdade e $\overline{B}$ denota seu complemento, ou seja, o conjunto de exemplos para os quais o corpo da regra é falso; similarmente para H e $\overline{H}$. Dessa forma, HB denota o conjunto de exemplos para os quais a cabeça da regra é verdade e o corpo da regra é também verdade, e assim por diante.

[3] Para maiores detalhes veja [6].

Tabela 1. Tabela de Contingência para uma Regra

	hb = número de exemplos para os quais H é verdade e B é verdade
	$\overline{h}b$ = número de exemplos para os quais H é falso e B é verdade
H $\overline{H}$	$h\overline{b}$ = número de exemplos para os quais H é verdade e B é falso
B hb $\overline{h}b$ b	$\overline{h}\overline{b}$ = número de exemplos para os quais H é falso e B é falso
$\overline{B}$ $h\overline{b}$ $\overline{h}\overline{b}$ $\overline{b}$	b = número de exemplos para os quais B é verdade
h $\overline{h}$ n	$\overline{b}$ = número de exemplos para os quais B é falso
	h = número de exemplos para os quais H é verdade
	$\overline{h}$ = número de exemplos para os quais H é falso
	n = número total de exemplos

Seja x a cardinalidade do conjunto X, ou seja, $x = |X|$. Na Tabela 1, h denota a cardinalidade do conjunto H, ou seja, $h = |H|$. Em outras palavras, h denota o número de exemplos para os quais a cabeça da regra é verdade. Similarmente $b = |B|$ denota o número de exemplos para os quais o corpo da regra é verdade, ou seja, o número de exemplos cobertos pelo corpo da regra. O número de exemplos para os quais a cabeça é verdade e o corpo é verdade é denotado por $hb = |HB|$.

Em problemas de classificação, hb representa o número de exemplos corretamente classificados pela regra e $\overline{h}b$ representa o número de exemplos incorretamente classificados pela regra. Da mesma forma, $h\overline{b}$ representa o número de exemplos não cobertos que são da classe da regra e $\overline{h}\overline{b}$ representa o número de exemplos não cobertos que não são da classe da regra.

Associada à cardinalidade x, a freqüência relativa f_x é então utilizada como uma estimativa da probabilidade $P(X)$, ou seja, $P(X) = f_x = \frac{x}{n}$. Por exemplo, considerando a cardinalidade hb, a probabilidade $P(HB)$ pode ser determinada da seguinte forma $P(HB) = f_{bh} = \frac{hb}{n}$. De forma semelhante podem ser determinados os valores das probabilidades $P(\overline{H}B), P(\overline{H}\overline{B})$ e $P(H\overline{B})$. Conhecidas essas probabilidades, os valores de $P(B)$, $P(\overline{B})$, $P(H)$ e $P(\overline{H})$ podem ser determinados.

As frequências relativas derivadas da tabela de contingência são utilizada para calcular diversas medidas de avaliação de regras encontradas na literatura. Essa medidas foram unificadas segundo o *framework* proposto por Lavrac em [1] como, por exemplo, a medida de cobertura de uma regra (Equação 1). Maiores detalhes sobre medidas de avaliação e de interessabilidade de regras veja [6].

$$covR = \frac{b}{n} = \frac{hb + \overline{h}b}{n} \tag{1}$$

No MAR do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ a medida apresentada na Equação 1, é usada para identificar se uma regra é ou não um pequeno disjunto. Nessa identificação, o usuário do conhecimento escolhe um valor pequeno para a medida de cobertura, e assim, todas as regras que apresentarem cobertura menor ou igual a esse valor são consideradas pequenos disjuntos. Em seguida, é procedido o cálculo do grau de surpresa.

2.1 Grau de Surpresa de Pequenos Disjuntos

Segundo Freitas [2], um pequeno disjunto é considerado como conhecimento surpresa quando esse disjunto prediz uma classe diferente das classes preditas pelas suas generalizações mínimas (GM). Um pequeno disjunto terá tantas GM's quantas forem as suas condições. Por exemplo, seja a regra *if-then*, descrita a seguir, um pequeno disjunto:

$$\textbf{if } cond_1 \textbf{ and } cond_2 \textbf{ and } cond_3 \textbf{ and } \cdots \ cond_m \textbf{ then classe } C_v$$

Cada uma das g generalizações está associada com uma das m condições do pequeno disjunto original ($g = 1 \cdots m$). Assim, as possíveis generalizações são:

1 **if** $cond_2$ **and** $cond_3$ **and** $\cdots$ $cond_m$ **then classe** y_1 (`removendo` $cond_1$)
2 **if** $cond_1$ **and** $cond_3$ **and** $\cdots$ $cond_m$ **then classe** y_2 (`removendo` $cond_2$)
3 **if** $cond_1$ **and** $cond_2$ **and** $\cdots$ $cond_m$ **then classe** y_3 (`removendo` $cond_3$)
$\cdots$
m **if** $cond_1$ **and** $cond_2$ **and** $\cdots$ $cond_{m-1}$ **then classe** y_m (`removendo` $cond_m$)

A classe y_g é a classe predita pela g-ésima generalização, ou seja, a classe de maior freqüência nos exemplos cobertos pela generalização g. A classe C_v é a classe do pequeno disjunto original. Deve ser observado que $\{y_{g=1\cdots m}, C_v\} \in \{C_1, C_2, \cdots C_{NCl}\}$.

O grau de surpresa do pequeno disjunto é o somatório do número de vezes que y_g é diferente de C_v, isto é:

$$SurpDisj(R) = \sum_{g=1}^{m} diferente\ (y_g, C_v) \tag{2}$$

onde

$$diferente = \begin{cases} 1 \text{ se } y_g \neq C_v \\ 0 \text{ caso contrário} \end{cases}$$

O resultado é um número inteiro no intervalo $0 \cdots m$. Quanto maior esse valor mais surpresa o pequeno disjunto apresenta, ou seja, mais interessante é o conhecimento por ele descrito. Em geral, pequenos disjuntos com grau de surpresa $= 0$ são desprezados.

Em virtude da identificação de pequenos disjuntos ser dada em função de valores pequenos da medida de cobertura, essas regras apresentam uma baixa generalização. Além desse aspecto, o usuário desse conhecimento também pode estar interessado em observar algum aspecto relacionado a precisão de classificação de exemplos. Para isso, podem ser usadas medidas de avaliação de precisão de regras, como as apresentadas a seguir.

2.2 Medidas de Precisão de Regras Utilizadas

Medidas de avaliação de regras são úteis para determinar quais são as "melhores" regras R_u de um classificador simbólico $\mathbf{h}$. Ainda não existe uma metodologia definitiva que verse a respeito da utilização dessas medidas, deixando a

critério do usuário do conhecimento escolher aquela(s) medida(s) que possam atender melhor suas expectativas.

Na construção do experimento aqui apresentado, os valores absolutos (cardinalidades) $[hb, \overline{h}b, h\overline{b}, \overline{h}\overline{b}, N]$ foram evidenciados para melhor compreender os valores calculados para as seguintes medidas de avaliação de precisão de uma regra:

- Precisão:

$$accR = P(H|B) = \frac{hb}{b} = \frac{hb}{hb + \overline{h}b} \tag{3}$$

  A medida de precisão $accR$ relaciona o número de exemplos corretamente classificados hb dentre o número de exemplos b cobertos pela regra. Toda vez que $accR$ é igual a 1 o valor absoluto $\overline{h}b$ é igual a 0. Essa medida tende a favorecer a precisão de exemplos positivos e, dessa forma, mesmo quando apresenta valor máximo ($accR = 1$) pode estar escondendo o fato de que o número de exemplos positivos (corretamente classificados) hb é pequeno em relação número total de exemplos n. Assim sendo, é mais indicada como uma medida de consistência da regra.

- Precisão de Laplace:

$$laccR = \frac{hb + 1}{b + NCl} = \frac{hb + 1}{hb + \overline{h}b + NCl} \tag{4}$$

  Foi incorporada a medida de *precisão de Laplace* [7] ao MAR para a realização do experimento aqui apresentado. Essa medida tem o propósito de melhorar a precisão $accR$ no sentindo de não privilegiar regras onde o número de exemplos corretamente classificados hb é pequeno. Diferentes regras com mesmos valores para hb e $\overline{h}b$ apresentam valores iguais para essa medida.

- Precisão Relativa com Peso:

$$wraccR = P(B)(P(H|B) - P(H)) = \frac{b}{n}\left(\frac{hb}{b} - \frac{h}{n}\right) \tag{5}$$

  A medida de precisão relativa com peso $wraccR$ tem o propósito de promover um balanceamento entre a generalidade e a precisão de uma regra. Diferentes regras com mesmos valores para hb, $\overline{h}b$ e $h\overline{b}$ apresentam valores iguais para essa medida.

3 O Experimento

Foram utilizados dois diferentes conjuntos de dados no experimento aqui apresentado foram obtidos da UCI [8].

O primeiro conjunto de dados utilizado foi o *Zoo Database* que contém informações, apresentadas na Tabela 2, para identificar 7 classes diferentes de animais.

Tabela 2. Informações do Conjunto de Dados *Zoo*

Nomes dos Atributos (Valores Possíveis)				
hair (y,n)	feathers (y,n)	eggs (y,n)	milk(y,n)	
airborne (y,n)	aquatic (y,n)	predator (y,n)		
toothed (y,n)	breathes (y,n)	venomous (y,n)		
fins (y,n) legs (y,n)	domestic (y,n)	tail (y,n) catsize (y,n)		
# Total de Exem.	Classes	# Exem.	Classe %	Erro CM
101	1	40	39,60%	60,40%
	2	20	19,80%	
	3	5	4,95%	
	4	13	12,87%	
	5	4	3,96%	
	6	8	7,92%	
	7	10	9,91%	

As informações apresentadas na Tabela 3 são referentes ao segundo conjunto de dados utilizado, *1984 United States Congressional Voting Records Database*, aqui chamado de *Voting*. Essas informação são utilizadas para identificar um candidato ao congresso norte-americano como democrata ou republicano.

Tabela 3. Informações do Conjunto de Dados *Voting*

Nomes dos Atributos (Valores Possíveis)				
water-project-cost-sharing (y,n)	handicapped-infants (y,n)			
adoption-of-the-budget-resolution (y,n)	mx-missile (y,n)			
physician-fee-freeze (y,n)	el-salvador-aid (y,n),			
religious-groups-in-schools (y,n)	anti-satellite-test-ban (y,n)			
aid-to-nicaraguan-contras (y,n)	immigration (y,n)			
synfuels-corporation-cutback (y,n)	education-spending (y,n)			
superfund-right-to-sue (y,n), crime (y,n)	duty-free-exports (y,n)			
export-administration-act-south-africa (y,n)				
# Total de Exem.	Classes	# Exem.	Classe %	Erro CM
434	democrat	266	61,29%	38,71
	republican	168	38,71%	

A Tabela 4 sumariza algumas informacões do terceiro conjunto de dados utilizado, *Car Evaluation Database*, aqui chamando de *Car*. O atributo *classe* refere-se a satisfação de consumidores de carros que pode ser não-aceitável (un-acc), aceitável (acc), boa (good) ou muito boa (v-good).

Utilizando o sistema de AM $\mathcal{CN}2$ [9], implementado em *C++* e disponível na biblioteca $\mathcal{MLC}$++ [10], foi induzido um classificador simbólico a partir de cada conjunto de dados anteriormente apresentado.

No experimento aqui realizado, os dados de entrada para o MAR do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ foram:

Tabela 4. Informações do Conjunto de Dados *Car*

# Exem.	Nome do Atributo	Valores Possíveis Atributo	Classe	Classe %	Erro CM
1728	buying	v-high, high, med, low	unacc	70.023%	29.97%
	maint	v-high, high, med, low	acc	22.222%	sobre unacc
	doors	2, 3, 4, 5-more	good	3.993%	
	persons	2, 4, more	v-good	3.762%	
	lug_boot	small, med, big			
	safety	low, med, high			

1. um conjunto de exemplos; e
2. um conjunto de regras induzidas.

Assim, cada conjunto de dados foi utilizado tanto para induzir quanto para avaliar o seu respectivo conjunto de regras.

Na Tabela 5 são apresentados os valores[4] das medidas de precisão, grau de surpresa e valores absolutos dos pequenos disjuntos identificados a partir dos conjuntos de dados *Zoo* e *Voting*. As regras foram enumeradas de acordo com a ordem em que foram induzidas.

Tabela 5. Medidas de Precisão e Surpresa dos Pequenos Disjuntos obtidos a partir dos Conjuntos de Dados *Zoo* e *Voting*.

# da	Medidas de Precisão				Surpresa	Valores Absolutos
Regra	`accR`	`wraccR`	`laccR`	`covR`	`SDisj`	$[hb, \overline{h}b, h\overline{b}, \overline{hb}, n]$
Zoo Database						
3	1.0000	0.0282	0.4000	0.0297	3.0000	[3,0,2,96,101]
4	1.0000	0.0188	0.3333	0.0198	3.0000	[2,0,3,96,101]
6	1.0000	0.0380	0.4545	0.0396	1.0000	[4,0,0,97,101]
7	1.0000	0.0729	0.6000	0.0792	1.0000	[8,0,0,93,101]
8	1.0000	0.0714	0.6000	0.0792	1.0000	[8,0,2,91,101]
9	1.0000	0.0357	0.4545	0.0396	2.0000	[4,0,6,91,101]
Voting						
4	0.7143	0.0016	0.6667	0.0161	1.0000	[5,2,262,166,435]
18	0.8750	0.0090	0.8000	0.0184	1.0000	[7,1,161,266,435]
19	0.8889	0.0104	0.8182	0.0207	1.0000	[8,1,160,266,435]

A partir do conjunto de dados *Zoo* foram induzidas 9 regras. Dentre estas regras foram descobertos 6 pequenos disjuntos que cobriam menos que 10% de exemplos, todos com grau de surpresa $\neq 0$. A partir do conjunto de dados *Voting* foram induzidas 21 regras, 4 pequenos disjuntos que cobriam menos que 10% de

[4] Esses valores tem a parte inteira separada da parte decimal por um ponto.

exemplos foram identificados e, destes, apenas 3 apresentaram grau de surpresa $\neq 0$.

Na Tabela 3 são apresentados os valores das medidas de precisão, grau de surpresa e valores absolutos dos pequenos disjuntos identificados a partir do conjunto de dados *Car*. As regras também foram enumeradas de acordo com a ordem em que foram induzidas. Os valores apresentados estão dispostos em ordem decrescente do grau de surpresa dos pequenos disjuntos.

Tabela 6: Medidas de Precisão e Surpresa dos Pequenos Disjuntos obtidos a partir do Conjunto de Dados *Car*

# da	Medidas de Precisão				Surpresa	Valores Absolutos
Regra	`accR`	`wraccR`	`laccR`	`covR`	`SDisj`[5]	$[hb, \overline{h}b, h\overline{b}, \overline{h}\overline{b}, n]$
67	0.7500	0.0012	0.5000	0.0023	4.0000	[3,1,381,1343,1728]
69	0.6667	0.0008	0.4286	0.0017	4.0000	[2,1,382,1343,1728]
96	0.6667	0.0011	0.4286	0.0017	4.0000	[2,1,67,1658,1728]
62	0.7500	0.0012	0.5000	0.0023	3.0000	[3,1,381,1343,1728]
63	0.7500	0.0012	0.5000	0.0023	3.0000	[3,1,381,1343,1728]
70	0.6667	0.0008	0.4286	0.0017	3.0000	[2,1,382,1343,1728]
75	1.0000	0.0005	0.4000	0.0006	3.0000	[1,0,383,1344,1728]
76	1.0000	0.0005	0.4000	0.0006	3.0000	[1,0,383,1344,1728]
88	0.7500	0.0016	0.5000	0.0023	3.0000	[3,1,66,1658,1728]
91	0.7500	0.0016	0.5000	0.0023	3.0000	[3,1,66,1658,1728]
92	0.6667	0.0011	0.4286	0.0017	3.0000	[2,1,67,1658,1728]
98	1.0000	0.0006	0.4000	0.0006	3.0000	[1,0,68,1659,1728]
101	1.0000	0.0006	0.4000	0.0006	3.0000	[1,0,68,1659,1728]
103	1.0000	0.0006	0.4000	0.0006	3.0000	[1,0,68,1659,1728]
112	0.7500	0.0016	0.5000	0.0023	3.0000	[3,1,62,1662,1728]
115	0.7500	0.0016	0.5000	0.0023	3.0000	[3,1,62,1662,1728]
18	1.0000	0.0002	0.4000	0.0006	2.0000	[1,0,1209,518,1728]
38	1.0000	0.0018	0.6250	0.0023	2.0000	[4,0,380,1344,1728]
48	1.0000	0.0018	0.6250	0.0023	2.0000	[4,0,380,1344,1728]
54	1.0000	0.0018	0.6250	0.0023	2.0000	[4,0,380,1344,1728]
58	1.0000	0.0018	0.6250	0.0023	2.0000	[4,0,380,1344,1728]
60	0.7500	0.0012	0.5000	0.0023	2.0000	[3,1,381,1343,1728]
61	0.7500	0.0012	0.5000	0.0023	2.0000	[3,1,381,1343,1728]
64	0.7500	0.0012	0.5000	0.0023	2.0000	[3,1,381,1343,1728]
66	0.7500	0.0012	0.5000	0.0023	2.0000	[3,1,381,1343,1728]
73	1.0000	0.0005	0.4000	0.0006	2.0000	[1,0,383,1344,1728]
74	1.0000	0.0005	0.4000	0.0006	2.0000	[1,0,383,1344,1728]
77	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]
78	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]
79	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]
81	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]
82	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]

continuaçãona próxima página

[5] Threshold = 0.0060 ou seja 0.6%

continuação da página anterior						
85	1.0000	0.0022	0.6250	0.0023	2.0000	[4,0,65,1659,1728]
86	0.7500	0.0016	0.5000	0.0023	2.0000	[3,1,66,1658,1728]
87	0.7500	0.0016	0.5000	0.0023	2.0000	[3,1,66,1658,1728]
89	0.7500	0.0016	0.5000	0.0023	2.0000	[3,1,66,1658,1728]
90	0.7500	0.0016	0.5000	0.0023	2.0000	[3,1,66,1658,1728]
93	1.0000	0.0006	0.4000	0.0006	2.0000	[1,0,68,1659,1728]
94	0.0000	0.0000	0.2500	0.0000	2.0000	[0,0,69,1659,1728]
95	1.0000	0.0006	0.4000	0.0006	2.0000	[1,0,68,1659,1728]
99	1.0000	0.0006	0.4000	0.0006	2.0000	[1,0,68,1659,1728]
102	1.0000	0.0006	0.4000	0.0006	2.0000	[1,0,68,1659,1728]
104	0.0000	0.0000	0.2500	0.0000	2.0000	[0,0,69,1659,1728]
113	0.7500	0.0016	0.5000	0.0023	2.0000	[3,1,62,1662,1728]
114	0.0000	0.0000	0.2500	0.0000	2.0000	[0,0,65,1663,1728]
45	1.0000	0.0018	0.6250	0.0023	1.0000	[4,0,380,1344,1728]
46	1.0000	0.0018	0.6250	0.0023	1.0000	[4,0,380,1344,1728]
55	1.0000	0.0018	0.6250	0.0023	1.0000	[4,0,380,1344,1728]
56	1.0000	0.0018	0.6250	0.0023	1.0000	[4,0,380,1344,1728]
59	1.0000	0.0018	0.6250	0.0023	1.0000	[4,0,380,1344,1728]
65	0.7500	0.0012	0.5000	0.0023	1.0000	[3,1,381,1343,1728]
68	0.7500	0.0012	0.5000	0.0023	1.0000	[3,1,381,1343,1728]
71	0.7500	0.0012	0.5000	0.0023	1.0000	[3,1,381,1343,1728]
72	0.7500	0.0012	0.5000	0.0023	1.0000	[3,1,381,1343,1728]
80	1.0000	0.0022	0.6250	0.0023	1.0000	[4,0,65,1659,1728]
83	1.0000	0.0022	0.6250	0.0023	1.0000	[4,0,65,1659,1728]
84	1.0000	0.0022	0.6250	0.0023	1.0000	[4,0,65,1659,1728]
97	0.0000	0.0000	0.2500	0.0000	1.0000	[0,0,69,1659,1728]
100	1.0000	0.0006	0.4000	0.0006	1.0000	[1,0,68,1659,1728]

Utilizando o conjunto de dados em Car foram induzidas 119 regras. Dentre estas regras, foram identificados 79 pequenos disjuntos, 59 com grau de surpresa $\neq 0$, que cobriam menos que 0.6% de exemplos.

4 Resultados Obtidos

Considerando os valores apresentados nas Tabelas 5 e 3 é possivel verificar que, mesmo pequeno disjuntos com valores iguais para o grau de surpresa podem apresentar diferentes valores para as medidas de precisão. No caso de pequenos disjuntos, os valores para $\overline{h}\overline{b}$ estão bem mais próximos de n que os valores hb, $\overline{h}b$ e $h\overline{b}$.

Nenhum pequeno disjunto apresentou $\overline{h}b \geq hb$. Quanto mais próximo o valor $\overline{h}b$ de hb, menor será o valor da medida $accR$. É interessante observar que o valor $h\overline{b}$ não interfere no valor obtido para $accR$, como pode ser visto, por exemplo, nos valores obtidos para as regras 7 e 8 do conjunto de dados *Zoo* ou também para as regras 69 e 92 do conjunto de dados *Car*. É possível verificar a medida de precisão $accR = 1$ para todos os pequenos disjuntos com $\overline{h}b = 0$, independente do

valor hb obtido, enfatizando que essa é uma medida de consistência do pequeno disjunto.

Observando os valores calculados para $laccR$, quanto maior o valor de hb e mais próximo de 0 for o valor de $\overline{h}b$, maior será o valor dessa medida para o pequeno disjunto. Diferentes regras com mesmos valores para hb e $\overline{h}b$ apresentam, também, valores iguais para essa medida, ou seja, da mesma forma de $accR$, o valor $h\overline{b}$ não interfere no valor obtido para $laccR$.

No caso das medidas $accR$ e $laccR$, as cardinalidades que tem relação com essas medidas são hb e $\overline{h}b$. Como os valores de $h\overline{b}$ não interferem nos resultados obtidos, é possível perceber que essas medidas não levam em conta o número de exemplos não cobertos que são da classe da regra. Assim, é possível dizer que essas medidas não levam em conta o número de exemplos que o disjunto deixou de cobrir que são da sua classe.

Quanto aos valores calculados para a medida $wraccR$, é possível verificar que existe um relação entre os valores hb, $\overline{h}b$ e $h\overline{b}$ e essa medida. Quanto maior o valor de hb e, ao mesmo tempo, menor o valor de $\overline{h}b$ e $h\overline{b}$, maior será o valor dessa medida. Dessa forma, é possível dizer que $wraccR$ apresenta menores valores para disjuntos que cobrem exemplos de uma outra classe e, ao mesmo tempo, aqueles que deixam de cobrir exemplos que são da sua classe.

5 Considerações Finais

Em tarefas voltadas para a descoberta de conhecimento, utilizando sistemas de AM supervisionado, deseja-se que as regras descobertas sejam mais genéricas e precisas tanto quanto possível. Sob o ponto de vista da precisão de classificação, pequenos disjuntos são regras que cobrem um número pequeno de exemplos e, em geral, são indesejáveis porque apresentam pouca generalidade e são propensos ao erro. No entanto, sob o ponto de vista do grau de surpresa, eles tem o potencial para capturar conhecimento inesperado a partir dos dados.

O usuário do conhecimento descrito por pequenos disjuntos pode estar interessado em observar, além do grau de surpresa, algum aspecto relacionado a cobertura e a precisão de classificação de exemplos. Este trabalho apresenta um experimento onde pequenos disjuntos, que apresentam grau de surpresa do conhecimento que descrevem, são avaliados utilizando diferentes medidas de precisão de regras.

Na análise dos valores obtidos, é possivel caracterizar uma relação direta entre cardinalidades e as medidas de precisão de regras utilizadas. Partindo dessa relação, é possível interpretar os valores obtidos para cada medida, sob o foco da cobertura e da precisão de classificação de exemplos. Espera-se assim compreender melhor o papel destas medidas na avaliação de pequenos disjuntos.

Deseja-se que, a partir da compreensão da relação entre os valores obtidos para cardinalidades e as medidas de avaliação e de interessabilidade, seja possível não só compreender melhor questões relacionadas com a cobertura e a classificação de exemplos por parte de pequenos disjuntos mas, num passo seguinte,

entender melhor o papel dessas medidas ao tratar essas mesmas questões em regras que cobrem um grande número de exemplos (grandes disjuntos).

Referências

1. Lavrac, N., Flach, P., Zupan, B.: Rule evaluation measures: a unifying view. In: Proceedings of the Ninth International Workshop on Inductive Logic Programming. LNAI. Volume 1634. (1999) 74–85
2. Freitas, A.A.: On objective measures of rule surprisingness. In: Principles of Data Mining & Knowledge Discovery: Proceedings of the Second European Symp. Lecture Notes in Artificial Intelligence. Volume 1510. (1998) 1–9
3. Bratko, I.: Prolog Programming for Artificial Intelligence. Addison-Wesley (1990)
4. Gomes, A.K., Monard, M.C.: Um módulo para avaliação de regras induzidas por algoritmos de aprendizado de máquina. In: Advances in Intelligent Systems and Robotics. (2003) LAPTEC-2003, Vol.II.
5. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Trans. on Knowledge and Data Eng. **8** (1996) 970–974
6. Gomes, A.K.: Análise do conhecimento extraído de classificadores simbólicos utilizando medidas de avaliação e de interessabilidade. (2002) Dissertação de Mestrado, ICMC-USP.
7. Clark, P., Boswell, R.: Rule induction with $\mathcal{CN}2$: Some recent improvements. In Kodratoff, Y., ed.: Proceedings of the 5th European Conference (EWSL 91). (1991) 151–163
8. Blake, C., Keogh, E., Merz, C.: UCI Repository of Machine Learning Databases (1998) http://www.ics.uci.edu/~mlearn/MLRepository.html.
9. Clark, P., Niblett, T.: The $\mathcal{CN}2$ induction algorithm. Machine Learning **3** (1989) 261–283
10. Kohavi, R., Sommerfield, D., Dougherty, J.: $\mathcal{MLC}$++: A Machine Learning Library in C++. IEEE Computer Society Press (1994)
11. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Los Altos, California, USA (1993)